INDIVIDUAL ARTICULATOR'S CONTRIBUTION TO PHONEME PRODUCTION

Jun Wang [1], Jordan R. Green [2], Ashok Samal [3]

[1]Callier Center for Communication Disorders
University of Texas at Dallas, Dallas, TX, United States
[2]MGH Institute of Health Professions, Boston, MA, United States
[3]Department of Computer Science & Engineering
University of Nebraska-Lincoln, Lincoln, NE, United States

ABSTRACT

Speech sounds are the result of coordinated movements of individual articulators. Understanding each articulator's role in speech is fundamental not only for understanding how speech is produced, but also for optimizing speech assessments and treatments. In this paper, we studied the individual contributions of six articulators, tongue tip, tongue blade, tongue body front, tongue body back, upper lip, and lower lip to phoneme classification. A total of 3,838 vowel and consonant production samples were collected from eleven native English speakers. The results of speech movement classification using a support vector machine indicated that the tongue encoded significantly more information than lips, and that the tongue tip may be the most important single articulator among all of the six for phoneme production. Furthermore, our results suggested that the tracking of four articulators (i.e., tongue tip, tongue body back, upper lip, and lower lip) may be sufficient for distinguishing major English phonemes based on articulatory movements.

Index Terms— Speech production, articulation, support vector machine, silent speech recognition

1. INTRODUCTION

Although most talkers produce speech effortlessly, the underlying coordination required to produce fluent speech is very complex involving dozens of muscles spanning the diaphragm to the lips. How exactly speech is produced is still poorly understood [1]. One major barrier to speech production research has been the logistic difficulty of tongue motion data collection [2]. Fortunately, recent advances in electromagnetic tracking devices have made speech production data collection more feasible. Tongue tracking using electromagnetic technology is accomplished through the placement of small sensors (or pellets) on the surface of the tongue. In prior work, the number of tongue sensors and their locations has been justified based on long-standing assumptions about tongue movement patterns, or the specific purpose of the study. It is, however, not clear how many sensors are adequate for a particular study because the individual articulator's contribution to the articulatory distinctiveness of phoneme production has rarely been studied.

Determining a minimal set of tongue sensors is important for optimizing (1) silent speech interface technologies designed to assist individuals with laryngectomy (surgical removal of larynx due to treatment of cancer) or severely impaired voice and speech [3, 4, 5, 6], (2) speech recognition with articulatory information [7, 8], and (3) treatments that provide a real-time visual feedback of speech movements [9, 10]. In addition, the use of more sensors than is necessary comes at a cost for both investigators and subjects; the procedure for attaching sensors to the tongue is time intensive and can cause discomfort and therefore, may limit the scope of research on persons with speech impairment.

In this research, we examined the individual contribution of six articulation points (articulators for the rest of the paper), tongue tip, tongue blade, tongue body front, tongue body back, upper lip, and lower lip to the articulatory distinctiveness of eight English vowels and eleven English consonants. Support vector machines (SVM, [11]) are a widely used machine learning classifier, which have been successfully used for classification of phonemes based on articulatory movements (e.g., [2, 12]). A SVM was used to classify vowel and consonant samples based on the movement of individual and groups of articulators. The resulting classification accuracies were used to address the following experimental questions:

Q1. *Which articulator contributes most to vowel production?*
Q2. *Which articulator contributes most to consonant production?*
Q3. *Is there a minimum set of articulators that can match the accuracy level achieved using all six articulators?*

2. DATA COLLECTION

2.1. Participants

Eleven native American English talkers participated in this study. No talker had positive history of speech or hearing problems. Each talker participated in one data collection session. Ten of them participated in a session for both vowels and consonants; the other one participated in a session for vowels only.

2.2. Stimuli

Eight major English vowels in consonant-vowel-consonant (CVC) form, /bɑb/, /bib/, /beb/, /bæb/, /bʌb/, /bɔb/, /bob/, /bub/, and eleven major English consonants in vowel-consonant-vowel (VCV) form, /ɑbɑ/, /ɑgɑ/, /ɑwɑ/, /ɑvɑ/, /ɑdɑ/, /ɑzɑ/, /ɑlɑ/, /ɑrɑ/, /ɑʒɑ/, /ɑdʒɑ/, /ɑjɑ/, were used as stimuli.

The eight vowels are representative of the full English vowel

set and were chosen because they sufficiently circumscribe the boundaries of articulatory vowel space [2, 13]. Each vowel was embedded in a consonant vowel consonant context. The pre and post vowel consonant was always /b/. This bilabial was selected because it is easy to parse and has minimum co-articulation effect on the vowel [2].

The eleven consonants were selected because they represent the primary places and manners of articulation of English consonants. Each consonant was embedded into the /ɑ/ context because this vowel is known to induce larger tongue movements than other vowels [2].

2.3. Procedure

The Electromagnetic Articulograph (EMA, Model: AG500; Carstens Medizintechnik, Inc., Germany) was used to register 3-D movements of the tongue, lip, and jaw during speech. The spatial accuracy of motion tracking using EMA (AG500) was 0.5 *mm* [14]. EMA registers movements by establishing a calibrated electromagnetic field that can be used to track the movements of small sensors within the field. The center of the magnetic field is the origin (zero point) of the EMA coordinate system.

Participants were seated with their head within the calibrated magnetic field. The sensors were attached to the surface of each tongue and jaw articulator using dental glue (PeriAcryl Oral Tissue Adhesive) and others using double-sided tape.

Figure 1 shows the placement of the twelve sensors attached to a participant's head, face, and tongue [2]. Three of the sensors were attached to a pair of plastic glasses. HC (Head Center) was on the bridge of the glasses; HL (Head Left) and HR (Head Right) were on the left and right outside edge of each lens, respectively. The movements of HC, HL, and HR sensors were used to calculate the movements of other articulators independent of the head [15]. Lip movements were captured by attaching two sensors to the vermilion borders of the upper (UL) and lower (LL) lips at midline. Four sensors - T1 (Tongue Tip), T2 (Tongue Blade), T3 (Tongue Body Front) and T4 (Tongue Body Back) - were attached approximately 10 *mm* from each other at the midline of the tongue [2, 15, 16]. The movements of three jaw sensors, JL (Jaw Left), JR (Jaw Right), and JC (Jaw Center), were recorded for future use, thus not analyzed in this study.

All stimuli were presented on a large computer screen in front of the participants and pre-recorded sounds were played to help the participants to pronounce the stimuli correctly. The stimuli were presented in the order as listed in Section 2.2. Participants were asked to repeat what they heard and to put stress on the middle phoneme (rather than the carriers) at their habitually comfortable speaking rate and loudness. Participants were also asked to rest shortly (about 0.5 second) between each CVC or VCV production to minimize the co-articulation effect [2]. This rest interval also facilitated segmenting the stimuli prior to analysis. Mispronunciations were rare, but were identified by the investigator and excluded from the data analysis.

All participants repeated the phoneme sequences multiple times. The sequences were then segmented into individual phoneme utterances offline, based on synchronously recorded acoustic data. On average, 21 valid vowel samples were collected from each participant with the number of samples for each vowel varying from 16 to 24 per participant. In total, 1704 vowel samples with 213 samples for each vowel were obtained. The average number of valid consonant samples collected from each participant was 19 varying from 12 to 24 per participant. In total, 2134 consonants samples (with 194 samples for each consonant) were obtained. In all, 3,838 vowel and consonant samples were collected and used for analysis.

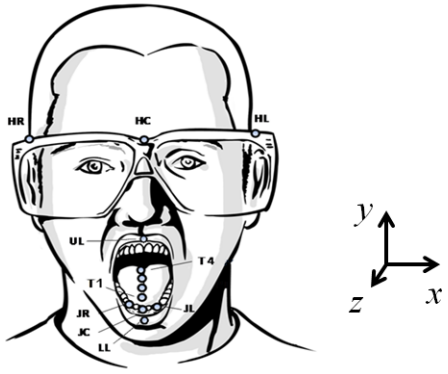

Figure 1: Sensor positions in data collection and the orientation of the Cartesian coordinate system. Sensor labels are described in text.

2.4. Data preprocessing

Prior to analysis, the translational and rotational components of head movement were subtracted from the tongue and lip movements. The resulting head-independent tongue and lower lip sensor positions included the movement from the jaw. The orientation of the derived 3-D Cartesian coordinate system is displayed in Figure 1. Because the movements for the simple vowels and consonants contain only very low frequency components, a low pass filter of 10 Hz was applied to the movement traces prior to analysis [15].

Only *y* (vertical) and *z* (anterior-posterior) coordinates of the sensors (i.e., T1, T2, T3, T4, UL, and LL) were used for analysis because the movement along the *x* (lateral) axis is not significant during speech of healthy talkers [16].

3. METHOD

Support vector machine [11] was used to classify those phoneme production samples based on the movement time-series from the six individual articulators, and for all possible combinations of those articulators.

SVM classifiers project training data into a higher dimensional space and then separate classes using a linear separator [11]. The linear separator maximizes the margin between groups of training data through an optimization procedure. Those training samples on the boundaries of the classes are called support vectors. A kernel function is used to describe the distance between two samples (i.e., *u* and *v* in Equation 1). The following radial basis function was used as the kernel function K_{RBF} in this study, where λ is an empirical parameter:

$$K_{RBF}(u,v) = \exp(1 - \lambda \parallel u - v \parallel) \quad (1)$$

For more details, please refer to [17], which describes the implementation of the SVM used in this study.

The same approach for constructing data samples in [2, 4, 5] was used in this study, where a sample (e.g., *u* or *v* in Equation 1) is a concatenation of time-sampled motion paths of articulators as

data attributes. Initially, the movement data of each individual articulator for each stimulus (a vowel or consonant) were time-normalized and sampled to a fixed length (i.e., 10 frames). The length was fixed, because SVM requires the input samples to be fixed-width array. The predominant frequency of tongue and lip movements is about 2 to 3 *Hz* for simple CVC or VCV utterances [18], thus 10 samples adequately preserve the motion patterns. Then, the arrays of y or z coordinates for those articulators were mean-normalized and concatenated into one sample (vector) representing a vowel or consonant. Overall, each sample contained $20 \times p$ (10 frames × 2 dimensions × p articulators) attributes for p articulators ($1 \leq p \leq 6$). An integer (e.g., 1 for /bɑb/, and 2 for /bib/) was used for labeling the training data.

Cross validation is a standard procedure for evaluating the performance of classification algorithms in machine learning, where training data and testing data are unique. In this study, Leave-*N*-out cross validation was used, where N (= 8 or 11) is the number of vowels or consonants, respectively. In each execution, one sample for each phoneme (totally N phonemes) in the dataset was selected for testing and the rest were used for training. There were a total of m executions; where m is the number of samples per phoneme. The average classification accuracy of all m executions was considered as the overall classification accuracy [19].

4. RESULTS AND DISCUSSION

4.1. Vowel classification on individual articulators

Figure 2 gives the average vowel classification accuracies across participants for each individual articulator. Paired-sample t-test showed the accuracy obtained from any single tongue articulator (i.e., T1, T2, T3, or T4) was significantly higher than that from UL or LL; the accuracy obtained from LL was significantly higher than that for UL ($p < 0.01$); there was no significant difference among the different tongue articulators. This finding might be explained by the tight biomechanical coupling between adjacent tongue regions [15].

In general, the findings suggested that tongue sensors contribute more to vowel classification than do the lips, a finding which is consistent with the long-standing descriptive knowledge in classical phonetics, in which vowels are distinguished by tongue height and front-back position [13]. The finding that the accuracy obtained from LL is higher than that for UL was not surprising, because the movement of LL included the movements of the jaw, which was a major articulator for vowel production [1, 15, 20].

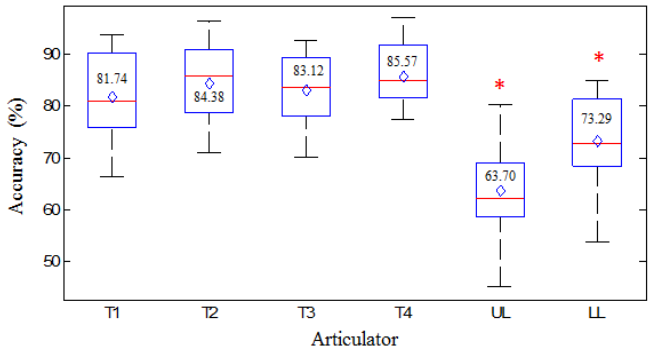

Figure 2. Average vowel classification accuracies across participants for individual articulators (diamond is the mean value; red line is the median; edges of the boxes are 25 and 75 percentiles).

4.2. Consonant classification on individual articulators

Figure 3 gives the average consonant classification accuracies across participants for each individual articulator. Similarly to the results for vowel classification and not surprisingly, accuracy obtained from any single tongue articulator was significantly higher than that for LL or UL, except T3 had no significant difference with LL; accuracy for LL was significantly higher than that for UL ($p < 0.01$). More interestingly, unlike the vowel classification results, the consonant classification accuracy obtained from T1 was significantly higher than that from T2 ($p < 0.05$), but no significant difference with that from T3 or T4. There was no significant difference observed among T2, T3, or T4.

The finding that T1 (Tongue Tip) contributes significantly more than T2 may reflect the quasi-independent movement of these regions during consonant production. When compared to the vowel findings, these findings suggested that consonant production involves more features (including place and manner of articulation), and that the tongue tip plays an important role in encoding these features. For example, dental consonants (e.g., /θ/) require tongue tip to have contact with teeth; and alveolar consonants (e.g., /l/) are produced with short distances between the tongue tip and alveolar ridge. Based on these findings, T1 appears to be the best sensor to use if only one tongue articulator can be used in a study.

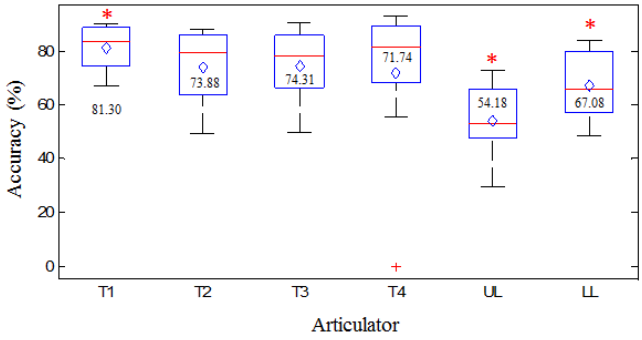

Figure 3. Average consonant classification accuracies across participants for each individual articulator (diamond is the mean value; red line is the median; edges of the boxes are 25 and 75 percentiles).

4.3. Classification on articulator combinations

To determine a minimum set of sensors that can be used to accurately classify speech movements, we compared the classification accuracies of all relevant combinations of articulators. We hypothesized that using only four articulators {T1, T4, UL, LL} that combined can capture the major movements of tongue and lips during speech. Our hypothesis was also informed by the observations reported in sections 4.1 and 4.2: T1 contributes significantly more in consonant production than T2 does; {T1, T4} obtained higher accuracy than {T1} or {T4}). Thus, Q3 in the end of Section 1 can be further refined as

Q4. *Is {T1, T4, UL, LL} a minimum set of articulators that can match the accuracy level achieved using all six articulators (i.e., {T1, T2, T3, T4, UL, LL})?*

To address this question we compared the classification accuracies of all relevant combinations of articulators. For the convenience of explanation, we name the hypothesized optimal combination/set

$$\boldsymbol{A} = \{T1, T4, UL, LL\} \quad (2)$$

First, the accuracy obtained from ***A*** was compared to those from combinations with fewer articulators (i.e., {T1, T4}, {T1, T4, UL}, and {T1, T4, LL}, and single articulators, {T1}, {T4}, {UL}, and {LL}) to verify no combination with fewer articulator than ***A*** has similar or higher accuracies than that for ***A***. Second, ***A*** was compared to those combinations without lip articulators but with more tongue articulators (i.e., {T1, T4, T2}, {T1, T4, T3}, {T1, T4, T2, T3}) to verify lip articulators are needed to avoid accuracy decrease. Finally, ***A*** was compared to those combinations with extra articulators (i.e., $\boldsymbol{A} \cup \{T2\}$, $\boldsymbol{A} \cup \{T3\}$, and $\boldsymbol{A} \cup \{T2, T3\}$) to verify that extra (tongue) articulators do not help to improve the classification accuracy.

Table 1 lists the accuracies obtained from ***A*** and from all other relevant combinations, as well as the significances between ***A*** and every other combination. As anticipated, the accuracy obtained from ***A*** was significantly higher than accuracy obtained from any combination with fewer articulators or any combination with extra tongue articulators but without lip articulators, which suggested that classification accuracy will decrease if all articulators in ***A*** are not included. Moreover, the addition of extra articulator on top of ***A*** did not increase the classification accuracy. Therefore, our results suggested {T1, T4, UL, LL} is a minimum set that can accurately encode articulatory distinctiveness of vowels and consonants.

Table 1. Average vowel and consonant classification accuracies across participants on selected articulator (sensor) combinations.

Articulator (Sensor) Combinations	Vowel Classification Accuracy (%)	Consonant Classification Accuracy (%)
{T1}	**81.74** ***	**81.30** ***
{T4}	**85.57** ***	**71.74** ***
{UL}	**63.10** ***	**43.18** ***
{LL}	**73.29** ***	**67.18** ***
{T1, T4}	**88.08** ***	**87.72** **
{T1, T4, UL}	**90.62** *	**89.97**
{T1, T4, LL}	**90.76** *	**90.10** *
{T1, T4, T2}	**86.88** ***	**89.97**
{T1, T4, T3}	**86.58** ***	**90.10** *
{T1, T4, T2, T3}	**85.70** *	**87.04** *
{**T1, T4, UL, LL**}	**91.65**	**91.36**
{**T1, T4, UL, LL,** T2}	**91.00**	**90.67**
{**T1, T4, UL, LL,** T3}	**90.87**	**90.85**
{**T1, T4, UL, LL,** T2, T3}	**90.02**	**90.85**

Significant differences between ***A*** ({**T1, T4, UL, LL**}) and every other combination are marked: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Relation to prior work. Although studies on speech articulation have often used three or four tongue sensors [2, 4, 5, 8, 15, 16, 20, 21, 22, 23, 24, 25], investigators have not empirically determined that this number of sensors is necessary. Our previous work [2] investigated the articulatory distinctiveness of vowels and consonants based on all six articulators, but not on individuals. Qin and colleagues [26] showed that three to four sensors are able to predict the tongue contour with only 0.3-0.2 *mm* error per point on the tongue surface. Those studies, however, did not reveal if fewer tongue articulators are sufficient for studies typically using three or four tongue sensors. To our best knowledge, this study is the first to empirically determine the optimal number of sensors and their locations for speech articulation studies. Of course, as mentioned previously, the number of sensors and their locations may vary depending on the purpose of the study and its application. For example, when investigating disordered speech articulation, it may be practical to use only two tongue sensors (typically tongue tip and tongue body back) [19, 20]. A single sensor (typically tongue tip) may also be adequate for treatment studies as well (e.g., [9, 10]).

5. CONCLUSION AND FUTURE WORK

This research studied the contribution of six articulators (i.e., tongue tip, tongue blade, tongue body front, tongue body back, upper lip, and lower lip, named as T1, T2, T3, T4, UL, and LL, respectively) to the production of major English vowels and consonants. A support vector machine was used to classify those vowel and consonant samples based on the movement of both individual articulators and their various combinations. The results indicated that any single tongue articulator had significantly higher contribution to both vowel and consonant production than did either lip articulator. Among the tongue articulators, T1 had significantly higher contribution than did T2 for consonant production, but no significant differences were observed among the other tongue articulators. In addition, our findings suggested {T1, T4, UL, LL} may be sufficient for typical assessment and treatment studies (e.g., a silent speech recognizer from articulatory movements), and that, if only one tongue articulator can be used, T1 conveys the most articulatory information.

Future work includes (1) extending the stimuli from phonemes to words and sentences, because the individual articulators may have different levels of contribution in word or sentence production, and (2) determining if the current findings are applicable to vowel and consonant production by talkers with motor speech disorders.

6. ACKNOWLEDGMENTS

This work was in part funded by Excellence in Education Fund, University of Texas at Dallas, Barkley Trust, University of Nebraska-Lincoln, and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States). We would like to thank Dr. Tom Carrell, Dr. Lori Synhorst, Dr. Mili Kuruvilla, Cynthia Didion, Rebecca Hoesing, Kate Lippincott, Kayanne Hamling, and Kelly Veys for their contribution to participant recruitment, data collection, and data processing.

7. REFERENCES

[1] R. D. Kent, S. G. Adams, and G. S. Tuner, "Models of speech production," in *Principles of Experimental Phonetics*, N. J. Lass, Ed., St Louis, MO: Mosby, 1996.

[2] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, 2013 (In press).

[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270-287, 2010.

[4] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces," *Proc. ICASSP*, pp. 4985-4988, 2012.

[5] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Whole-word recognition from articulatory movements for silent speech interfaces," *Proc. Interspeech*, 2012.

[6] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering & Physics*, vol. 30, no. 4, 419-425, 2008.

[7] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of Acoustical Society of America*, vol. 121, no. 2, 723-742, 2007.

[8] F. Rudzicz, G. Hirst, P. Van Lieshout, Vocal tract representation in the recognition of cerebral palsied speech, *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 4, 1190-1207, 2012.

[9] J. Levitt and W. F. Katz, "The effects of EMA-based augmented visual feedback on English speakers' acquisition of the Japanese flap: A perceptual study," *Proc. Interspeech*, pp. 1862-1865, Makuhari, Japan, 2011.

[10] W. F. Katz and M. McNeil, "Studies of articulatory feedback treatment for apraxia of speech (AOS) based on electromagnetic articulography," *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, vol. 20, no. 3, pp. 73-80, 2010.

[11] C. Cortes and V. Vapnik, "Support-vector network," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[12] J. Wang, J. R. Green, A. Samal, and D. B. Marx, "Quantifying articulatory distinctiveness of vowels", *Proc. Interspeech*, pp. 277-280, Florence, Italy, 2011.

[13] P. Ladefoged and K. Johnson, *A course in phonetics* (6th Ed.). Independence, KY: Cengage Learning, 2011.

[14] Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for AG500 electromagnetic articulograph," *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 547-555, 2009.

[15] J. R. Green and Y. Wang, "Tongue-surface movement patterns during speech and swallowing," *Journal of Acoustical Society of America*, vol. 113, pp. 2820-2833, 2003.

[16] J. Westbury, *X-ray microbeam speech production database user's handbook*, University of Wisconsin, 1994.

[17] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.

[18] J. R. Green, E. M. Wilson, Y. Wang, and C. A. Moore, "Estimating mandibular motion based on chin surface targets during speech," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 928-939, 2007.

[19] J. Wang, *Silent speech recognition from articulatory motion*, Doctoral dissertation, University of Nebraska-Lincoln, 2011.

[20] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 596-611, 2008.

[21] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," *Phonus*, vol. 5, pp. 1-13, 2000.

[22] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523-541, 2011.

[23] F. H. Guenther, C. Y. Espy-Wilson, S. E. Boyce, M. L. Matthies, M. Zandipour, and J. Perkell, "Articulatory tradeoffs reduce acoustic variability during American English /r/ production," *Journal of Acoustical Society of America*, vol. 105, pp. 2854–2865, 1999.

[24] J. S. Perkell, F. H. Guenther, H. Lane, M. L. Matthies, E. Stockmann, M. Tiede, and M. Zandipour, M. "The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts," *Journal of Acoustical Society of America*, vol. 116, pp. 2338–2344, 2004.

[25] M. Stone, M. Epstein, K. Iskarous, "Functional segments in tongue movement," *Clinical Linguistics & Phonetics*, vol. 18, no. 6-8, pp. 507-521, 2004.

[26] C. Qin, M. A. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," *Proc. Interspeech*, pp. 2306-2309, 2008.